ARCHIVAL DATA PROJECTS

Jean E. Crampon
Science & Engineering Library
University of Southern California[1]

ABSTRACT: This will be a report on two archival data projects undertaken at the Hancock Library of Biology and Oceanography of the University of Southern California to improve access to both textual and visual archival data owned by the Library. The first project was to digitize unpublished station data for use in a relational database. The second project was a prototype of a system for accessing photographs and negatives. The methodologies for selecting, designing, executing, and evaluating these two projects will be discussed.

PROJECT 1 STATION DATA

Background

Within the Hancock Library of Biology & Oceanography's archival files are the original station data for the voyages of research vessels sponsored by the University of Southern California. The data used for this project were collected during voyages of the *Velero IV* from 1948 to 1983 and include standard parameters such as station number, latitude and longitude, location, wind direction and speed, depth, equipment used, specimens collected, etc. (See Figure 1 Sample Station Data Sheet.) Researchers who use these data are often working with specimens collected during the voyages. The station number is part of the identification of the specimen and is marked on the specimen itself, a tag on the specimen or the specimen container for microscopic specimens. The number allows the related data to be retrieved. This level of detail is not available from any other source than the station data sheets. Significant collections of specimens from these voyages are permanently housed at many institutions including the California Academy of Sciences, the Los Angeles County Museum of Natural History, the Santa Barbara Museum of Natural History, the Smithsonian, and others. Researchers from all over the world have consulted or borrowed specimens over the years.

Subsets of the data have appeared in various publications of the Allan Hancock Foundation, but the information in the total data set, or simply a single specific station entry, is only available by direct request to the Hancock Library. Although some of the researchers who use these data have requested improved access, nothing had been done to make available all the data collected by the *Velero IV*. In the 1980's there was an attempt to microfilm the data sheets, but this was not technically feasible due to the condition of

---

[1] Formerly Head Librarian, Hancock Library of Biology and Oceanography, University of Southern California.

the originals. Many of the data sheets are much too light to be microfilmed clearly. In the summer of 1995 the Library was given $2000 through the College of Letters, Arts, and Sciences (LAS) of the University to be used within one year to support a student to enter the data into a database and make the data not only more accessible, but to allow more modes of access than simply by date or station number.

Execution of the Project

There was no money allotted for additional equipment or software. We wanted to use a relatively basic program that would be available over time and would be easy for the researcher who would use it later. After some investigation of what was available, Microsoft Access® was selected for the relational database program. Since this was the first time I had used this database software, I was able to obtain some assistance in setting up the data entry form to make it easy to enter the data and be assured that no data elements were being omitted. A random sampling of the original data sheets was examined for data elements needed to create the template. (See Table 1 Data Elements for the Station Data Project.)

Table 1
Data Elements for the Station Data Project

Cruise Number
Recorders
Date of Cruise (MM/DD/YY)
Time Zone
Station Number
General Locality
Gear Used
Starting Latitude, Longitude, Time and Depth in fathoms
Ending Latitude, Longitude, Time and Depth in fathoms
Direction of Travel
Wind Direction and Speed
Remarks

Over the next year two students in succession were hired for this project. Their job was to do the data entry. Data entry was difficult for many reasons. The terminology was unfamiliar to the students and the original data recorders varied in their skills. Sometimes the recorder was a scientist, but often they were members of the ship's crew. Problems included handwriting, varieties of spelling or abbreviation of place names, samples, and equipment, and fading of original sheets over time. (See Figure 2 Sample Data Entry Record.) When the students had a question they recorded it on a log for review. (See Figure 3 Sample Log Page.) I soon learned that it was necessary to deal with the questions from the log quickly, but I also learned I needed to proofread all of the entries. Quick response to questions limited the number of repetitive errors. Proofreading was necessary for two reasons: 1. The original data sheets were difficult to read and 2. The mind-numbing nature of data entry in itself caused multiple errors.

The first student hired worked during the summer of 1995 for about twenty hours a week. When she resigned a second student was selected to work on the project. This student worked for ten months, only nine of which was paid from the original funding. Because all the original funding had not been spent within the original twelve-month period, I applied to continue over the summer of 1996. LAS granted this, so the entire time was a total of fifteen months. There was not active data entry for all of the months because data entry was suspended while undergoing the searches for the student assistants. After the funding ceased we were able to hire the second student as a regular Hancock Library student, but she was not able to include very many hours on this project in addition to standard student responsibilities. (See Table 2.)

Table 2 Project Hours

Time Period	Student Hours	Librarian Hours	Funding Source
FY96 Q1	79	20	LAS
FY96 Q2	126.5	16	LAS
FY96 Q3	105.3	17	LAS
FY96 Q4	73.9	46	LAS
FY97 Q1	77.5	9	LAS
FY97 Q2	4	8	Library
FY97 Q3	0	0	Library
FY97 Q4	0	3	Library
FY98 Q1	0	4	Library
FY98 Q2	0	0	Library
FY98 Q3	0	6	Library
Total Hours	466.2	129	

What did I learn from this project?

First, data entry was much slower than was predicted before the project began. Approximately 40% of the data were entered during the entire funding period. Second, it took many more hours of my own time than originally thought. The proofreading was extremely time consuming. The time expenditure was approximately four hours of student time to one hour of mine. I learned how far behind in proofreading the students' work I could get without causing a scheduling problem for myself. Third, I had to coordinate my schedule carefully with the student. During the funding period the software was only available on one machine in the library and that machine was on my desk, so I had to vacate my office for them to do their work efficiently. Fourth, the selection of students was very important. The first student (Student A) only worked during the first summer, but I hired her because she came highly recommended and had data entry experience. Unfortunately, her data entry experience was all with numerical data entry. The second student (Student B) turned out to be better at data entry where the data are not all numerical. Differences between the students included background (Student A was an engineering student, Student B was pre-nursing so had

taken more science courses), keyboarding skills (Student B typed faster than Student A), and personality (Student A was less willing to ask questions than Student B, which may have been a cultural difference). Language difficulty was less understanding of English than of the scientific terminology used. Neither student was a native English speaker. Fifth, I needed an advocate to get the original funding. LAS gave the original $2000 at the request of one of the research faculty. Since there then was an approved project, I was in position to request the extension in time. This extension kept the project going for an additional three months with no increase in the total funding committed.

Currently the project is at a standstill. There is an opportunity for additional funding; however, some of the "cast of characters" has changed. I will be continuing to work on the project, but there are new people in the coordinating roles for the administration and LAS. The requesters remain the same. The people still highly interested are the original requesters, including the requester who lobbied for the original funding. The completion of this project looks highly probable with additional funding to support it. The expanded access to the data could make this attractive to be used in environmental research to be able to compare current conditions with conditions at these locations up to fifty years ago.

Results

The project is not completed, but an excellent start has been made. Due to the size of the data set, the original plan to make the data available on disk has changed somewhat. This has been driven by changes in technology as well. The present data set, and the entire data set when completed, will be available in the Microsoft Access® format as designed; however, the size of the data set has caused a change in the archiving media from multiple floppy disks to a single zip or CD-ROM disk. The original requesters want the data to be published in print form as well although the expanded searching capabilities of the software would be unavailable in a hard copy format. Publication will depend on the extent of future funding although the first priority is digital access.

PROJECT 2 PHOTO ARCHIVE

Background

The second project was a joint project that was the result of a class. In fall of 1993 the USC Center for Software Engineering under the leadership of Prof. Barry Boehm experimented with a master's level class to focus on skills that students would need in the real world software engineering environment. (Boehm et al. 1998) Although the end result would certainly require programming and design, the course uses the WinWin® (Horowitz 1996) theory and software tool to help the students understand roles of the software architect, the developer, the customer, the user, etc. in a software development project. (See Table 3.) From 1993 through 1996 each student on the team was assigned one of the "stakeholder" roles although all were to participate in the programming

Table 3
What is WinWin?

WinWin is a computer program that aids in the capture, negotiation, and coordination of requirements for a large system. It assumes that a group of people, called *stakeholders*, have signed on with the express purpose of discussing and refining the requirements of their proposed system. The system can be of any type. WinWin contains facilities for:

1. capturing the desires (win conditions) of the stakeholders
2. organizing the terminology so that stakeholders are using the same terms in the same way
3. expressing disagreements or issues needing resolution
4. offering options as potential solutions
5. negotiating agreements which resolve the issues
6. using third party tools to enlighten or resolve issues
7. producing a requirements document that summarizes the current state of the proposed system
8. creating documents that support multimedia and hyperlinks
9. tracing the ways by which requirements decisions were reached
10. checking the completeness and consistency of requirements.

and writing of documentation. The application of the WinWin tool encouraged the refining process. Prof. Boehm found that the students tended not to ask the questions that a "real" customer or user would ask. In Fall 1996, Prof. Boehm worked with two library faculty (Julie Kwan and Denise Bedford) to recruit projects with a library faculty member as the "client" for the students. It was also felt the library faculty might have a better understanding of what the "user" would need. For the fall 1996 class the library faculty members interested in participating prepared a brief abstract to apply to work with the students. (See Table 4 Hancock Photo Archive.)

Table 4
Problem Set: Hancock Photo Archive

There is a substantial collection of photographs, slides, and films in some of the Library's archival collections. As an example of the type of materials available, I would like to suggest using the archival collections of the Hancock Library of Biology and Oceanography to see if better access could be designed. Material from this collection is used by both scholars on campus and worldwide. Most of the Hancock materials are still under copyright, but the copyright is owned by USC in most cases.

For each abstract the two library faculty, along with Prof. Boehm, determined general feasibility before offering the abstracts to the students for their selection. In fall 1996 there were fifteen teams of six students each. The teams were self-selected. Each team applied for assignment to one of twelve projects based on the abstracts submitted by the library faculty. Prof. Boehm approved the student assignments to a project. The Hancock Photo Archive project was one of three projects assigned to two student teams. Each team had to do a separate project without consultation across teams. Over the semester I met with each team separately and discussed the primary needs of the project, was presented with sample screens and programs and attended class presentations on the various projects. I also participated in the evaluation of the two teams that worked on my project. Results from one of the teams were selected for presentation to the Dean of the Library. This is the project that will be discussed here.

Execution of the Project

The final required document for the student projects is a *Life Cycle Architecture Package*. This is the most complete documentation for a project. Most of the discussion of the project is based on this documentation. The students named their project the Hancock Digital Multimedia Archive System or HDMA. I quote here from their documentation with additional comments from me in brackets:

> "The purpose of the Hancock Digital Multimedia Archive (HDMA) System is to replace the labor intensive, manual system of organizing and cataloging the photographic materials in the Hancock collection with a multimedia capable computer based system. This system will facilitate access to the information and will make the collection available to a wide variety of users. [The archive occupies approximately ten filing drawers, two shelves, and four racks of film canisters. The photos are in better order than either the films or the slides.]

The primary objectives of the system include:

1. To provide a means for the Hancock Library of Biology and Oceanography to organize and catalog the collection of photographs, films, and reports from the Hancock expeditions. [Reports were not an original part of this project, but reference to the reports with abstracts was intended to be included if the project was accepted for further development.]
2. To make the catalog and photographs available to students, researchers and the public via the Internet.
3. To create an electronic backup of the photographs in the collection to safeguard the information in the event of a disaster. [This was very important to the students.]

The Hancock collection consists of thousands of photographic slides, negatives, film and reports that were recorded during the various Hancock scientific expeditions of the

1930's [into the 1980's, but the bulk of the materials are from the *Velero III* voyages from 1931 until World War II, so their date is close]. The photographs are a priceless record of the research conducted during the voyages and were donated to the Hancock Library in the 1950's [actually donated sporadically over the entire time the photographs were taken]. The collection has never been properly cataloged, hence access to the collection's photographs is poor and a valuable source of research is underutilized.

The Hancock Digital Media Archive will provide the means for librarians, students and researchers to browse and/or query the collection, from anywhere in the Internet, in order to locate an item of interest. The system will provide a means to locate items based on various attributes such as subject, photographer, date, etc. In addition, the system will provide a digital image of the photograph of sufficient resolution to identify the subject. This image will contain a USC copyright notice in order to protect the rights of the University. A means to order a print of any desired photograph or report will be provided. [Ordering of reports was an enhancement the students strongly felt was important as this process is completely manual.]

Finally, the HDMA system will create a backup of the material in the collection. "In the event of a disastrous earthquake, or other calamity, the priceless record of the Hancock expeditions will be preserved." (*Life Cycle* 1996)

Negotiations between the team and me continued through the fall semester. Specific features or functions were agreed upon and the students developed a "proof-of-concept" prototype. The system was required to display text and static images with a thumbnail image and a larger image with a copyright statement superimposed on the image to protect it. (See Figure 4 Sample Archival Copyright Photograph.)

A means of requesting a copy of the image without the copyright statement was built into the system although credit card payment was not a feature as the University would not permit it. The system had to be easy to use, allow browsing the images as well as searching, use existing computing infrastructure of the University, be accessible through the Internet, and be able to adapt to the Library's database management system (SIRSI). A separate controlled password-only access by the librarians in Hancock would allow the capability to add, update, or delete records. The system had to "provide search results in 10 seconds 90% of the time [and] must be able to download a 30KB image file in 30 seconds 90% of the time." (*Life Cycle* 1996) It also would detect errors due to communications problems, servers, etc.; be expandable in the future to include dynamic (video/film) media; and provide ordering capabilities.

Additional features were considered but rejected in this phase, such as full text of reports and film digitization. The large amount of storage for digitizing film was not available to the project, but was considered as a future enhancement. Digital cash was also considered as a future enhancement when the University allowed it. With the current system for providing photographs from the collection, most of which still have not been digitized, an estimate of photographic reproduction was also not feasible at the time.

Features of the administrative program were limited to the Hancock Library. (See Table 5 Administrative Module Features.) The subject terms were free text and entered by the librarian in the administrative mode. Generally these were intended to be accessible by a non-scientist, so species were not always entered. The photo archive did not usually contain this information, so addition of this would require supplemental work by the librarian in conjunction with an appropriate scientist and was determined to be a future enhancement if later found necessary. Subject terms were easily edited by the librarian and some hierarchical group terms were added as needed, e.g. if the subject was dolphins, then marine mammals was used also. In the case of the prototype, the students sometimes used their own somewhat casual terminology, e.g. "bunch of dolphins at sea."

Table 5: Administrative Module Features

Edit Photograph Records	Generate Alphabetic Pages
List Photograph Records	Generate Report List
Edit Report Records	System Reports
List Report Records	Online Help
Edit Subject Word List	Go to the HDMA home page

Features of the Internet access mode would be available to anyone. (See Table 6 Internet Access Digital Multimedia Archive Features.) The searching access allowed the user to restrict a search by subject, date, or place. Place was a textual search, such as Guadalupe Island, not by longitude and latitude as that was generally not available in the photo archive. The user also could search photos, reports, or both. The default was photos.

Table 6: Internet Access Digital Multimedia Archive Features

Search Archive	View Order Form
Browse Photos	Process Order
Browse Reports	Online Help

The fall semester class (CS577a) is a core course in software engineering, but the spring semester (CS577b) is not. The second semester class is much smaller and similar projects were combined. Of the twelve projects from the fall, six were selected for continuation in the spring. This was based primarily on the interests of the continuing students and not necessarily the librarians' priorities. Of the membership of the two Hancock teams, only one student enrolled in the spring class. The project was combined with two other image-based projects with the lead taken by the project whose members were in the majority. Although it was not intended, major points of the two other projects, including the Hancock project, were lost in the spring results. (Boehm et al. 1998) This included the copyright notice and some of the searching options, such as photographer and date. Since these were class projects, University Computing automatically deleted all the projects without consultation with the faculty. As of spring of 1998 the projects were thought to be archived, but that was found not to be the case when I requested the project prototype. What is left is a large amount of written documentation, but none of the computer programs are viewable. Since all of the students have completed their degrees and left the

University, I have been unable to contact any of the students to see if they archived any data themselves.

What did I learn?

First, this was as much a learning process for the development of the course as a prototype development.(Mankin 1997) Due to the protocols then in place for class projects the prototype was not retained on the university's computer. This is no longer true. All projects are now to be retained for future revisiting.

Second, since the Hancock teams' students did not return for the second semester, the planned continuation of the project with enhancements did not occur. Combination of teams that had the same project is now acceptable for the spring semester, but not trying to combine different projects. All the image projects that year were not compatible. Too many elements of the individual projects were lost.

Third, there has to be significant "buy in" by the University Library for a project to continue. For the most current class year 1998-99 the project selection and support has begun to be institutionalized. New people in the R&D unit are involved and funding for some continuation is provided.

Fourth, I learned as much about the process of negotiation for a software engineering project as the students did. We each had to learn about the other's terminology.

Fifth, the time invested was worth it to me to bring focus to a needed function within the Hancock Library, i.e., better organization and access to non-text archival material. Access to original research data is a current interest within the University Library in its new identity. (See Results section below.) Because I worked with these archives, I have a better knowledge of their organization and arrangement. For example, the description of an image is often separate from the image, particularly for negatives.

Sixth, problems were identified in the technology. The materials include photographs, standard slides, glass slides, 4x4 negatives, and films. Treatment of each of these materials is different, so we started with the easiest, the photographs. Availability of a scanner was also a problem. The scanner available was best at photographs and either could not scan other formats or did not handle them well.

Seventh, the possibility of revisiting earlier projects is now an option. With the changes in the structure of the class the image project could be revived and the printed documentation in hand will be invaluable.

Results

Unfortunately for this project the only tangible result is the printed documentation. Essentially all computer programming would have to be redone, although the code was

submitted as part of the class. The class has been revised and possibilities of funding for future projects are now in place. Due to a reorganization on campus in fall 1997, there are closer relationships between the Library and University Computing. They are combined, along with Telecommunications and Student Information Services, into the current Information Services Division (ISD). The reorganization includes an R&D unit that is involved in selection of projects for the software engineering class. This should make development of future projects more viable. There is also an interest on the part of the head of ISD, the Chief Information Officer, in making original data available.

Conclusions

Both projects were worthwhile applications of my time and effort. Although neither has been fully developed, both have given visibility to the Hancock archival collections on campus that would be difficult to achieve in any other way. Possibilities for future work in these areas still exist. The increased interest in original data may work to the advantage of these projects or future projects in these areas. So far, on campus focus on original data has been on social sciences and humanities, but this science data archive is one that has good development potential.

References

Boehm, B. et al. 1998. Using the WinWin Spiral Model: A Case Study. *Computer* 31(7):33-44.

Horowitz, E. 1996. *WinWin Reference Manual. A System for Collaboration and Negotiation.*

Life Cycle Architecture Package: Hancock Digital Multimedia Archive. Team #10. CSCI-577a Fall 1996. USC, December 4, 1996.

Mankin, E. 1997. Library gives real-world challenge to student software design. *USC Chronicle* 16(18):7,12. [Online]. Available: http://www.usc.edu/dept/News_Service/chronicle/pdf/.

Figure 1
Sample Station Data Sheet

The University of Southern California

Allan Hancock Foundation

VELERO IV STATION LOG

ptain Allan Hancock, commanding — Recorded by

ıae			Month		Year									Time Zone
	62		April		1951									+8
tta. No.	Date Apr.	General Locality	Gear	Start Lat. N.	Start Long. W.	Start Time	Start Fms.	Finish Lat. N.	Finish Long. W.	Finish Time	Finish Fms.	Bottom	Remarks
210	13	Hassler Cove, San Martin Island, B. Cfa.		30-29	116-06-00	1100	shore			1200		Rock	Reef and Tidepools
212	13	Lagoon, San Martin I.		30-28-50	116-06-27	1230	shore			1330		Rock	Lagoon
213	14	Rocky Point San Quintin Bay		30-22-50	115-57-45	1030	shore			1130		Rock	Rocky outcrops in lagoon
214	14	Inside San Quintin Bay	boat dredge	30-22-50	115-57-30	1400	4½	30-24-35	115-57-10	1530	2½	Sand, Sh	Ulva, Spirorbis & Balanus, Broken Shell
215	14	Bahia San Quintin	night light	30-21-45	115-57-15							surface	
216	15	164° 13 miles from San Geronimo I.	Sigsbee	27-35-30	115-44-15	0900	48	28-34-15	115-43-00	0940	49	rocky	Torrey Pines, Lept. Diaptera, Chiton
217	15	16307 14.2 miles E of San Geronimo I. Baja Cal.	Biol dredge	28-34-15	115-43-00	0945	53	28-33-45	115-42-45	1005	54	Green mud	
218	15	Blanca Bay	Biol dredge	29-05-41	114-41-56	1700	12	29-05-47	114-41-23	1709	8	fine sand	
219	15	Blanca Bay	Biol dredge	29-05-39	114-41-15	1729	8	29-06-15	114-46-57	1734	6	Sand	sand dollars
220	16	Seammon Lagoon	boat dredge	27-50-45	114-17-40	1400	4	27-49-00	114-17-45	1500	6	sand	Dead shell
221	16	Shell Island, Scammon L.		27-49-30	114-15-20	1430	shore			1500	shore	sand	Pectens + gastropods
222	17	Point 10 mi. W of Pt. Malarrima		27-49-00	114-43-00	1000	shore			1230	shore	rock	Reef + Tidepools, Rich
223	17	" " " 4	lobster pots	27-48-03	114-40-08	1800	5	18 April		0600	5	rocky bottom	
224	18	9.5 miles West of Pt Malarrima	Biol dredge	27-49-33	114-43-30	0914	7	27-49-00	114-42-09	0922	9	rocks sand	
225	18	10 mi. W. of Pt. Malarrima		27-49-00	114-43-00	1130	shore			1330	shore	rock	Reef + Tidepools
226	19	So. Bay Cedros I. Mex	biol dredge	28 05 00	115-19-45	1035	16	28 04 49	115-19-24	1046	19	mud sand	
227	19	" " " " "		28-05-55	115-20-10	1200	shore			1400	shore	rock	Reef + Tidepools
228	20	3.5 miles 133° T from Mouth Redondo Pt Cedros	Biol Dredge	28-00-47	115 09 00	1035	33	28 01-09	115-08-42	1041	34	Sand	Solid Ophiothrix Mixed shells
229	20	2.6 miles 131° T from Mouth Redondo Pt Cedros	Sigsbee Trawl	28 00-49	115-08-56	1055	34	28 01 09	115-08 40	1115	35	Sand	Solid Ophiura + Ophiothrix
230	20	5.4 miles 093° T. from Mouth Redondo Pt Cedros	"	28 02 23	115-06-08	1148	40	28-02-27	115-05-36	1210	41	green mud	Coelenterata
231	20	N. end of Cedros Island		28-21-20	115-11-45	1430	—			1630	—		Land collecting Helix

Figure 2: Sample Data Entry Record

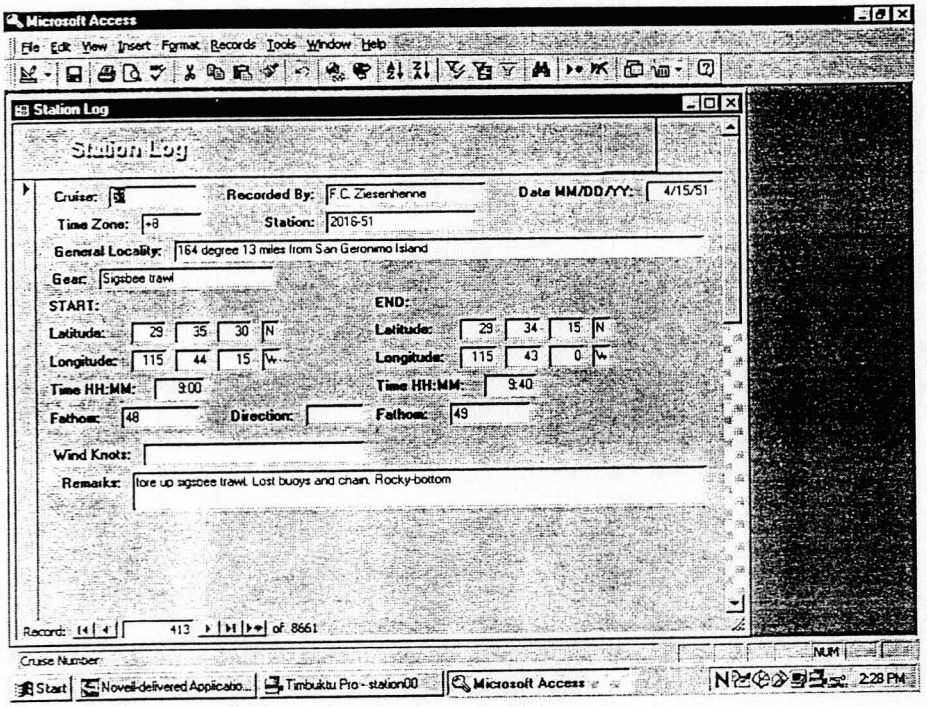

Figure 3: Sample Log Page

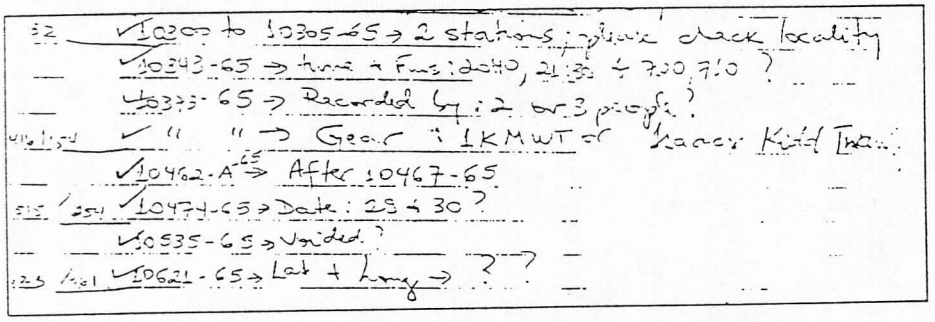

Figure 4: Sample Archival Copyright Photograph